Title

dbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals.

Journal

Nucleic acids research, 46(D1)

ISSN

0305-1048

Authors

Huang, Moli
Chen, Ye
Yang, Manqiu
et al.

Publication Date

2018

DOI

10.1093/nar/gkx796

Peer reviewed

Published online 5 September 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D71–D77***
doi: 10.1093/nar/gkx796

dbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals

Moli Huang[1,2,3,*], Ye Chen[2], Manqiu Yang[1], Anyuan Guo[4], Ying Xu[3], Liang Xu[2,*] and H. Phillip Koeffler[2,5,6]

[1]School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, China, [2]Cancer Science Institute of Singapore, National University of Singapore 117599, Singapore, [3]Cambridge-Suda Genomic Research Center, Soochow University, Suzhou 215123, China, [4]Department of Biomedical Engineering, Huazhong University of Science and Technology, Wuhan 430074, China, [5]Division of Hematology/Oncology, Cedars-Sinai Medical Center, University of California Los Angeles School of Medicine, Los Angeles, CA 90048, USA and [6]National University Cancer Institute, National University Hospital, 119074, Singapore

Received July 19, 2017; Revised August 17, 2017; Editorial Decision August 27, 2017; Accepted August 30, 2017

ABSTRACT

Core transcription regulatory circuitry (CRC) is comprised of a small group of self-regulated transcription factors (TFs) and their interconnected regulatory loops. Studies from embryonic stem cells and other cellular models have revealed the elementary roles of CRCs in transcriptional control of cell identity and cellular fate. Systematic identification and subsequent archiving of CRCs across diverse cell types and tissues are needed to explore both cell/tissue type-specific and disease-associated transcriptional networks. Here, we present a comprehensive and interactive database (dbCoRC, http://dbcorc.cam-su.org) of CRC models which are computationally inferred from mapping of super-enhancer and prediction of TF binding sites. The current version of dbCoRC contains CRC models for 188 human and 50 murine cell lines/tissue samples. In companion with CRC models, this database also provides: (i) super enhancer, typical enhancer, and H3K27ac landscape for individual samples, (ii) putative binding sites of each core TF across the super-enhancer regions within CRC and (iii) expression of each core TF in normal or cancer cells/tissues. The dbCoRC will serve as a valuable resource for the scientific community to explore transcriptional control and regulatory circuitries in biological processes related to, but not limited to lineage specification, tissue homeostasis and tumorigenesis.

INTRODUCTION

Transcription factors (TFs), also known as sequence-specific DNA-binding proteins, act typically through *cis*-regulatory elements to coordinate the transcription of target genes. While human and murine genomes harbor thousands of TFs (1), a small number of master TFs which are often expressed in a cell type/lineage-specific manner control disproportionately the transcriptional programs governing cell state and cell identity (2–6). Identification and characterization of core transcriptional regulatory networks are essential for better understanding of cell/tissue homeostasis and for addressing fundamental molecular and cellular biologic questions. Seminal studies from embryonic stem cells (ESCs) have revealed that core TFs, including NANOG, SOX2 and POU5F1/OCT4 dominate the transcriptional programs of pluripotency, self-renewal, and determination of cell fate (7–12). These core TFs not only bind to their own loci, but also regulate mutually, thereby forming cross-regulated feed-forward loops which maintain the pluripotency, but retain responsiveness to differentiation signals (9,13). The core TFs and their interconnected auto-regulatory loops have been termed 'core transcription regulatory circuitry' (CRC) (9).

In addition to ESCs, a growing body of literature demonstrates CRCs in other cellular systems and uncovers their fundamental roles in both cell type/lineage-specific and disease-promoting transcriptional programs. By using chromatin immunoprecipitation (ChIP) and high-resolution promoter microarrays, Odom *et al.* revealed a highly interconnected and auto-regulated transcriptional regulatory circuitry for six liver-enriched master TFs (HNF1α, HNF4α, FOXA2, ONECUT1, CREB1 and USF1) in human hepatocytes (14). By using ChIP-seq analysis, Sanda *et al.* identified that an oncogenic TF TAL1 forms a positive interconnected auto-regulatory loop with GATA3 and

*To whom correspondence should be addressed. Tel: +86 12 6588 0103; Fax: +86 12 6588 0103; Email: huangml@suda.edu.cn
Correspondence may also be addressed to Liang Xu. Tel: +65 6516 2162; Fax: +65 6873 9664; Email: csixl@nus.edu.sg

RUNX1 in multiple human T-cell acute lymphoblastic leukemia (T-ALL) cells, with MYB being a key downstream target (15). Of note, within TAL1-positive T-ALL cells (e.g. Jurkat), TAL1 enhancer is targeted recurrently by somatic small insertions which create new MYB binding sites and super enhancers (SEs) (5,16). The recruitment of MYB to TAL1 enhancer and subsequent identification of MYB-TAL1 complex further establish MYB as an extended CRC member for T-ALL cells with TAL1 enhancer mutation (16,17). In alveolar rhabdomyosarcoma, Gryder *et al.* uncovered that the fusion PAX3-FOXO1 is a TF which reprograms the myogenic SE landscape and establishes an oncogenic CRC together with three master TFs (MYCN, MYOD, and MYOG) (18). These experimentally identified interconnected auto-regulatory loops are generally critical to maintain the transcriptional homeostasis under both physiological and pathological conditions.

SEs are clustered enhancers densely bound by an array of TFs, and have been widely-studied to drive the expression of cell-type-specific and/or disease-associated genes, including core TFs (5,17–20). SEs are usually identified by the ROSE program using ChIP-seq signals of histone marks (H3K27ac and H3K4me), BRD4, mediators, or cell type-specific master transcription factors as input (20,21). Importantly, Saint-André *et al.* recently developed a method (CRC Mapper) to reconstruct CRC models based on the identification of SE-associated core TF genes in samples with SE maps. This approach can not only recapitulate and expand the experimentally verified CRCs, but also predict CRC models for samples of interest (17,22–24). For example, Aldiri *et al.* used CRC Mapper to reveal the involvement of dynamic CRCs during retinal development (22). By applying similar principles in SE landscapes of primary medulloblastoma tissues, Lin *et al.* discovered specific sets of TFs associated with subgroup identity of this disease and created a CRC blueprint for each subgroup (25). Based on the CRC model inferred from the BRD4 ChIP-seq data set in MOLT4 T-ALL cells, Winter *el al.* reported that core TF genes are very sensitive to chemically induced degradation of BET bromodomain proteins (26). Together, these studies demonstrate the successful CRC reconstruction to address critical research questions related to tissue development, cancer biology and drug responses, highlighting the broad and important utility of CRC modeling in biologic and medical investigations.

Here, we present dbCoRC (http://dbcorc.cam-su.org), the first comprehensive and interactive database of CRC models, with the aim to identify CRCs for diverse cell/tissue types and provide a valuable resource to fulfill the fast-expanding scientific exploration of transcriptional control and regulatory circuitries in normal and disease conditions. The dbCoRC bears the largest and most updated dataset of SEs in human (N = 188) and murine samples (N = 50). Unlike the dbSUPER (27) and SEA (28), two established databases centered on super enhancer cataloging, our database focuses on the construction of interconnected auto-regulated transcriptional loops, based on the research advance of core transcriptional regulatory circuitry (17). Currently, the dbCoRC contains and visualizes CRC models for over 230 samples, all of which are computationally inferred from H3K27ac ChIP-seq-defined SE landscape and the prediction of TF binding sites across SE regions (17). It also incorporates general descriptions of core TFs, together with their expression data in normal or cancer cells/tissues. To facilitate customized analysis or further data mining, this database supports data extraction of: (i) SE, typical enhancer, and H3K27ac ChIP-seq signals in individual samples and (ii) annotated binding motifs of each core TF across the SE regions within CRC. Hence, the dbCoRC will serve as a convenient platform to store, search, and analyze CRC-related data. These CRC models in the dbCoRC would be useful for the exploration of both cell/tissue type-specific and disease-associated transcriptional networks.

MATERIALS AND METHODS

Data sources

In the current study, we integrated the H3K27ac ChIP-seq data from the ENCODE Consortium (29), NIH Roadmap Epigenomics Mapping Consortium (30), and many other studies (Supplementary Table S1). Raw H3K27ac ChIP-seq data sets were downloaded through NCBI Gene Expression Omnibus (GEO) database. This version of dbCoRC provides CRC models for 188 human and 50 murine cell lines/tissue samples. Among these samples, 79 of them are cancer cells/tissues, and 159 of them are non-tumor cells/tissues (Table 1). Detailed information of each sample (source, tissue type, cell type, health status) can also be found in Supplementary Table S1.

Data processing and CRC reconstruction

Based on publicly available data sets of H3K27ac ChIP-seq assays, we reconstructed a CRC model for each sample consistent with a previously published method (17). Key steps for CRC modeling (Supplementary Figure S1) are summarized below.

Identification and mapping of SE. For each sample, raw H3K27ac ChIP-seq data sets (in .sra format) were firstly converted using the SRA toolkit. ChIP-seq reads were subsequently mapped using Bowtie1.1.2 (31) to either the human (hg19) or murine (mm9) reference genome with parameters -m 1 -k 1 –best. MACS 1.4.2 (32) was used for peak calling with parameters -p 1e-9. SEs were identified using ROSE (20) with parameters -t 2000. During enhancer stitching and ranking, H3K27ac peaks which occurred within ± 1 kb to a transcription start site (TSS) were subtracted. SEs were then assigned to the closest genes. When multiple closest genes were identified to be associated with same SE, this SE was assigned to a TF gene (if any) for subsequent analysis.

Predication of active SE-associated TFs. H3K27ac read counts within promoter region (± 1 kb to the TSS) of each gene/transcript were ranked in each sample. Transcripts with their promoter H3K27ac signals ranking top 2/3 were considered expressed actively. Next, 1,253 TFs were retrieved from the intersection of AnimalTFDB (1) and TcoF (33) databases, and subjected to the identification of SE-associated active TFs.

CRC modeling based on motif scanning in SE regions. To begin, 3160 DNA binding motifs for 695 TFs were compiled from the TRANSFAC database and MEME suite

Table 1. Data summary in the dbCoRC database

Species	Biosample	Number of samples (cancer/non-cancer)	Average number of TFs in CRC	Average number of SE-associated genes
Homo sapiens	Cell lines (67)	58/9	13	559
	Embryonic stem cells (2)	0/2	15	624
	Induced pluripotent stem cells (1)	0/1	10	412
	In vitro differentiated samples (14)	0/14	12	683
	Primary cells (32)	4/28	14	559
	Tissues (72)	16/56	17	725
Mus musculus	Cell lines (2)	1/1	16	548
	Embryonic fibroblasts (2)	0/2	17	729
	Embryonic stem cells (3)	0/3	18	913
	In vitro differentiated samples (5)	0/5	25	895
	Primary cells (3)	0/3	12	498
	Tissues (35)	0/35	16	661

TF: transcription factor; CRC: core transcriptional regulatory circuitries; SE: super-enhancer

(17,34,35). ROSE-defined SE regions were extended 500 bp on each side and followed by motif scanning with FIMO (36). Putative self-regulated TF was identified, if one SE-associated TF had at least three binding motifs within its own extended SE region. Within the same sample, motif scanning was applied further to identify potential binding sites of all auto-regulated TFs in their extended SE regions. All possible fully interconnected auto-regulatory circuitries were then constructed, scored and ranked in a given sample. For each candidate regulatory circuitry, its score is calculated by dividing the overall times of occurrence of core TFs across all possible circuitries by the number of core TFs in this circuitry. The top model which contained TFs with the highest frequency of occurrence across all possible circuitries was selected as the model of CRC in individual samples.

DATABASE FEATURES AND USE

Web interface and general functions

The dbCoRC provides a user-friendly easy web interface to browse, search, visualize and download. A top navigation bar is designed to assist individuals to use the above mentioned functions of this database, and access raw data/original publications related to each sample ('Data Sources'), general summary of this database ('Statistics'), as well as frequently asked questions ('Help').

The 'Home' page not only introduces CRC with both rolling pictures and text descriptions, but also offers a 'Quick search' function for straightforward inquiry. The search function also recognizes NCBI recorded gene alias (37). The 'Browse' page is organized alphanumerically as a sortable and interactive table which enables both fast fuzzy search for samples and customized filter through species, biosample types, and tissue/cell types. The number of records (10, 25, 50 and 100) per page can be increased/decreased using the 'show entries' dropdown menu. To view the CRC model for a given sample, users can simply click the sample name. Subsequently, an interactive CRC image showing the SE positions and interconnected regulations of core TFs will be displayed, together with corresponding tabular information of core TFs which can be either copied to the clipboard or exported as Excel and CSV files. Further clicking on a core TF of interest will open a new page showing: (i) its general description with links to external sources including NCBI Gene (37), Ensembl (38), OMIM (https://omim.org/), wikigenes (39), GEO and PubMed; (ii) its potential downstream targets and upstream regulators; and (iii) its expression in normal/tumor samples in the data sets extracted from TCGA (https://cancergenome.nih.gov/), EBI Expression Atlas (40), CCLE (http://www.broadinstitute.org/ccle) or RhesusBase (mouse TF genes) (41). The 'Search' page allows users to explore one or multiple potential core TF genes in this database. The 'Visualize' page provides links for users to visualize data in the UCSC genome browser (42) and GREAT server (43). The 'Download' page provides all downloadable files of each sample, including Zip files of core TF binding sites (.bed), SE annotation (.bed), processed H3K27ac ChIP-seq signals (.bw) and peak annotation (.xlsx). Users can easily select and batch download the data of interest.

A case application of the dbCoRC to explore the CRC in H1 human ESCs

From the 'Browse' page, users can apply three strategies to search samples of interest. To find 'H1 human ESCs' in this case (Figure 1A), users can (option 1) select 'human' as 'Species' and 'Embryonic Stem Cell' as 'Biosample type'; (option 2) enter 'H1' in 'Search' tab; or (option 3) adjust the number of entries per page and look up the alphanumerical table. After clicking the 'H1', the page will be redirected to display a visualized and interactive CRC model (Figure 1B) in companion with exportable tabular information of individual core TFs and respective SE regions (Figure 1C). The interactive image of the interconnected loops will respond to the movement of user's mouse. When mouse moves over the 'NANOG' position, the SE information of NANOG and all the mutual interactions between NANOG and other core TFs in this sample will be displayed (Figure 1B, lower panel). Further clicking the 'NANOG' either on the image or in the table will open a new page with four panels showing the general description of NANOG, potential downstream core TF targets of NANOG, potential upstream core TFs of NANOG, and the expression of NANOG in normal/tumor samples, respectively. The 'Description' panel shows the contents of

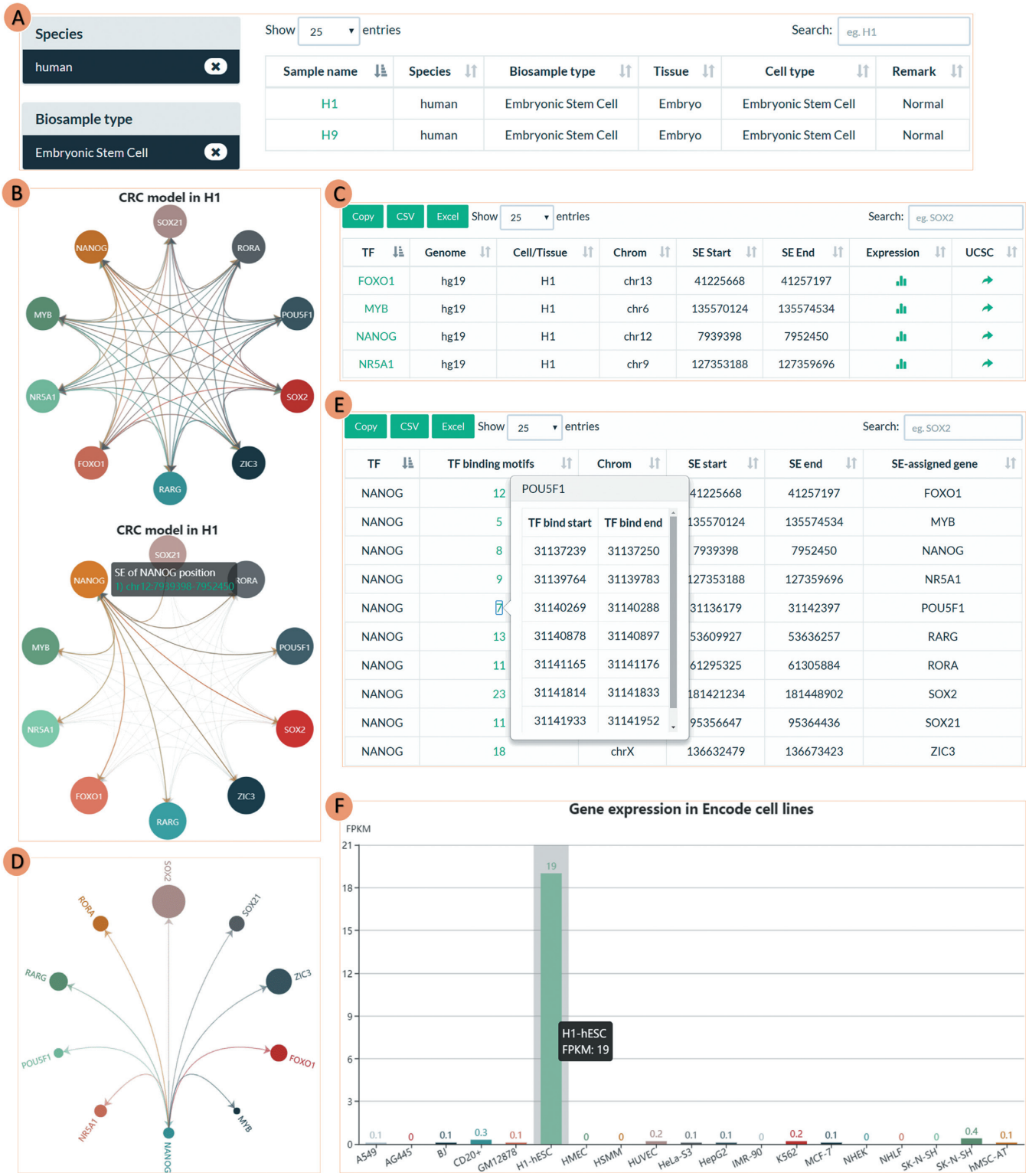

Figure 1. Interactive browsing in the dbCoRC. (**A**) Browse database for H1 human embryonic stem cells. Users can filter samples through species, biosample types, and tissue/cell types. A 'Search' box also enables fast fuzzy search. (**B**) Visualization of CRC model for H1 cells. (**C**) Exportable tabular information of core TFs in the H1 CRC model. (**D**) Potential downstream targets of NANOG within the H1 CRC model. (**E**) Interactive and exportable table showing the potential NANOG binding sites within the SE regions of core TFs in H1 cells. (**F**) Expression of NANOG mRNA across a panel of human cell lines.

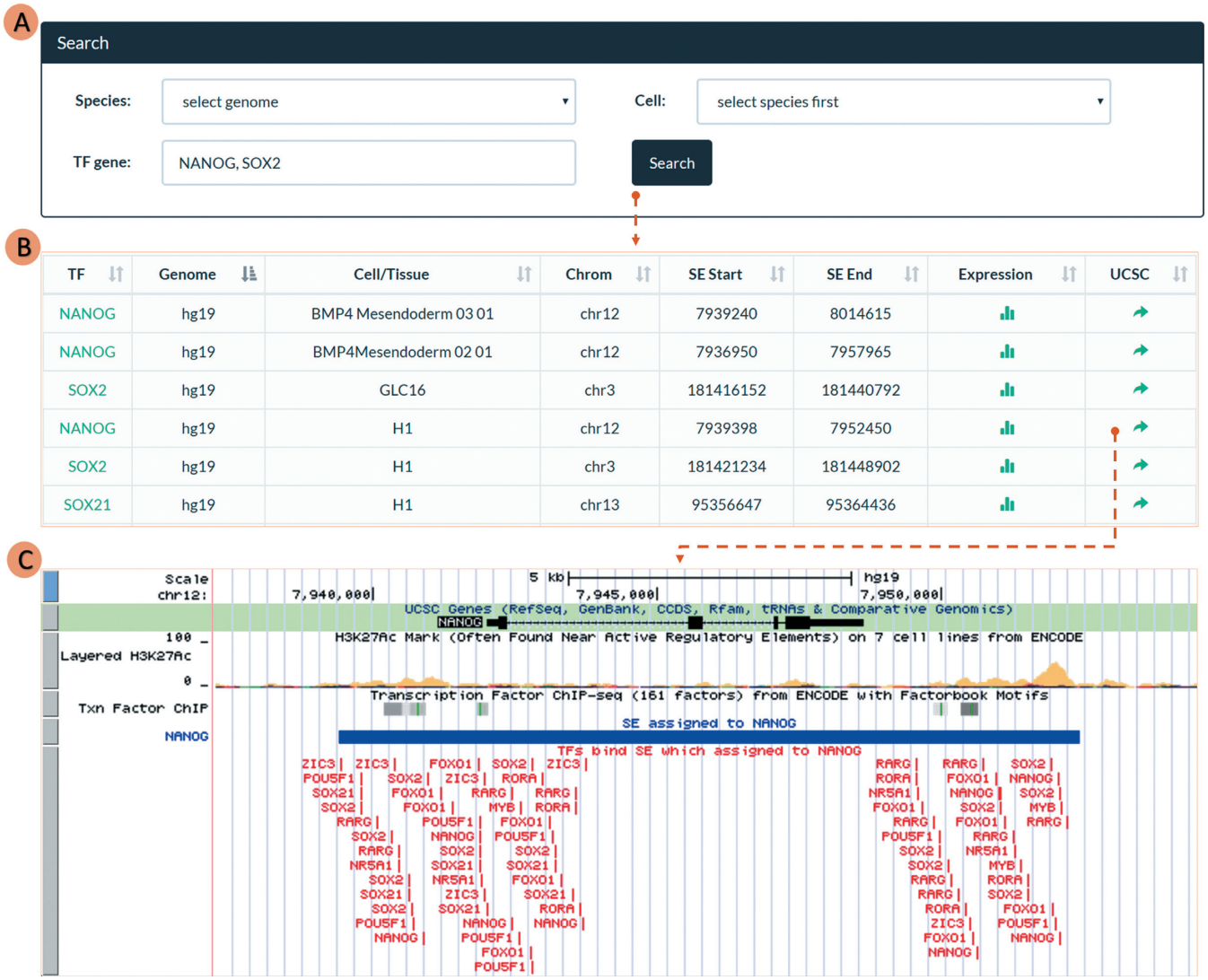

Figure 2. Searching in the dbCoRC and linking with external web servers for visualization. (**A**) Fuzzy search of NANOG and SOX2 in the database. (**B**) Results of search query from (A). (**C**) Visualization of the putative binding sites of core TFs across the extended SE region of NANOG via the UCSC genome browser.

gene ID, official symbol, full name, Ensembl ID, and RefSeq summary of NANOG, as well as the source and remarks of the H3K27ac ChIP-seq data in H1 cells. Users can also be linked to external databases including NCBI Gene, Ensembl, OMIM, and wikigenes for further information. The panels 'Potential downstream targets of NANOG within the CRC model' and 'Potential upstream regulators of NANOG within the CRC model' display the potential regulatory events between NANOG and other core TFs as both interactive images and tables (Figure 1D and E). The diameters of circles in the images of these two panels correspond to the numbers of TF binding sites within the SE region of a target gene. From the interactional tables, users can view, sort, and export the individual binding sites of a core TF of interested across the SE regions of another core TF gene. The 'Expression' panel demonstrates the differential expression of NANOG across a large panel of human cell lines, normal tissues, and cancers. Of note, NANOG is selectively expressed in H1 hESC compared with other cell lines (Figure 1F).

To query one or several TFs (e.g. NANOG and SOX2) in the dbCoRC, users can go to the 'Search' page, type 'NANOG, SOX2' into the 'TF gene' box, and click the 'Search' button (Figure 2A). Samples containing either NANOG or SOX2 in their CRCs will be returned as hits and organized as an interactive and sortable table (Figure 2B). Of note, SOX21 in H1 cells can also be identified as a result of a fuzzy search for 'SOX2'. Users can include further 'Species' or 'Cell' to restrict/refine the search. By clicking the TF in a sample of interest (NANOG in H1 cells), users will be guided to a new page with four panels showing its general information, regulation within H1 CRC, and expression (Figure 1D–F). Via the external link provided in the table, users can transfer directly data from this database to UCSC Genome Browser, and subsequently visualize the putative binding sites of core TFs across the NANOG SE

region in H1 cells (Figure 2C). If visualization and/or analysis of all core TF binding sites within H1 CRC is desired, the dbCoRC provides additional options in the 'Visualize' page with external links to UCSC Genome Browser and GREAT server. Moreover, users can search and download raw data and annotations of H1 H3K27ac ChIP-seq data (e.g. H3K27ac.bw) from the 'Download' page for in-house and/or in-depth bioinformatics analyses (Supplementary Figure S2).

Therefore, the dbCoRC provides a platform to store, search, visualize, and analyze CRCs. The information from this database may provide novel insights into the transcriptional regulation in a given sample, strongly encouraging follow-up biological and functional investigations.

SYSTEM DESIGN AND IMPLEMENTATION

The current version of the dbCoRC was constructed on the basis of MySQL and operated on an Ubuntu server. The interactive and responsive user interface was built with Bootstrap (http://getbootstrap.com) and JQuery (http://jquery.com/), which will resolve the compatibility issues across various devices, such as a laptop, smart phone or tablet. Visualization of data is implemented using the eCharts library (http://echarts.baidu.com/). The dbCoRC database is freely accessible to the scientific community via the web link (http://dbcorc.cam-su.org).

DISCUSSION AND FUTURE DEVELOPMENT

Understanding core transcriptional regulatory networks will help to address fundamental biologic questions related to cell/tissue type-specific and disease-associated transcriptional regulation. In this work, we developed dbCoRC, the first user-friendly and interactive database of CRC models. Compared with the dbSUPER (27) and SEA (28), two established super enhancer databases, the dbCoRC focuses on the construction of core transcriptional regulatory circuities in a large group of human and murine samples. To date, we have identified 330 core TFs which occur at least once in these 238 samples. We hope that our database will be helpful for the following users/researchers: (i) those interested in key transcriptional programs governing self-renewal and lineage specification/differentiation; (ii) scientists interested in key dysregulated transcription factors governing development and progression of disease; (iii) those interested in the conserved/core transcriptional modules regulating cell/tissue homeostasis; (iv) individuals who want to conduct a large-scale or in-depth in silico analysis of tissue/cell-type specific cis-regulatory elements, super enhancers or master transcription factors.

Continuous efforts will be made to update the CRC data and improve the functionality of this database. Current version of the dbCoRC is developed based on H3K27ac ChIP-seq data, because (i) ChIP-seq data are most available from H3K27ac than other SE-associated proteins, and (ii) H3K27ac can be used as a surrogate indicator for gene expression (17). Input data for CRC reconstitution from H3K27ac ChIP-seq will be extended to additional ChIP-seq data sets which are applicable for SE mapping. To improve the TF information (e.g. TF list and their DNA binding motifs) used for CRC modeling, additional TF resources, such like TFdb (http://genome.gsc.riken.jp/TFdb/), DBD (44), and TF ChIP-seq data will be integrated in future studies. The extended CRC network (17) will also be incorporated into this database.

Overall, the goal of the dbCoRC is to serve as a valuable resource for the scientific community to explore transcriptional regulation and genetic circuits in biologic processes related, but not limited to lineage specification, tissue homeostasis and tumor development.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We would like to thank Dr Yu Xue (Huazhong University of Science and Technology), Dr Takaomi Sanda (Cancer Science Institute of Singapore) and HPK lab members for the useful suggestions and kind help. We thank Richard A. Young for sharing CRCmapper program to this work.

FUNDING

Singapore Ministry of Education Academic Research Fund Tier 2 [MOE2017-T2-1-033 to H.P.K.]; Singapore Ministry of Health's National Medical Research Council (NMRC) Centre Grant awarded to National University Cancer Institute of Singapore; National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiatives; Natural Science Foundation of Suzhou [SYS201517 to M.H.]; Jiangsu Government Scholarship for Oversea Study; the Priority Academic Program Development of Jiangsu Higher Education Institutions; Innovation Research Fund for Graduate Students of Soochow University [2016xj061 to M.H.]; Science and Technology Bureau of Suzhou [MCMX201605 to Y.X.]. Funding for open access charge: National Research Foundation Singapore under its Singapore Translational Research (STaR) Investigator Award [NMRC/STaR/0021/2014 to H.P.K.].

Conflict of interest statement. None declared.

REFERENCES

1. Zhang,H.M., Liu,T., Liu,C.J., Song,S., Zhang,X., Liu,W., Jia,H., Xue,Y. and Guo,A.Y. (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76–D81.
2. Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
3. Graf,T. and Enver,T. (2009) Forcing cells to change lineages. *Nature*, **462**, 587–594.
4. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
5. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
6. Neph,S., Stergachis,A.B., Reynolds,A., Sandstrom,R., Borenstein,E. and Stamatoyannopoulos,J.A. (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.

7. Avilion,A.A., Nicolis,S.K., Pevny,L.H., Perez,L., Vivian,N. and Lovell-Badge,R. (2003) Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.*, **17**, 126–140.
8. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
9. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
10. Wang,Z., Oron,E., Nelson,B., Razis,S. and Ivanova,N. (2012) Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell*, **10**, 440–454.
11. Kim,J., Chu,J., Shen,X., Wang,J. and Orkin,S.H. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.
12. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
13. Chew,J.L., Loh,Y.H., Zhang,W., Chen,X., Tam,W.L., Yeap,L.S., Li,P., Ang,Y.S., Lim,B., Robson,P. *et al.* (2005) Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.*, **25**, 6031–6046.
14. Odom,D.T., Dowell,R.D., Jacobsen,E.S., Nekludova,L., Rolfe,P.A., Danford,T.W., Gifford,D.K., Fraenkel,E., Bell,G.I. and Young,R.A. (2006) Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.*, **2**, doi:10.1038/msb4100059.
15. Sanda,T., Lawton,L.N., Barrasa,M.I., Fan,Z.P., Kohlhammer,H., Gutierrez,A., Ma,W., Tatarek,J., Ahn,Y., Kelliher,M.A. *et al.* (2012) Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell*, **22**, 209–221.
16. Mansour,M.R., Abraham,B.J., Anders,L., Berezovskaya,A., Gutierrez,A., Durbin,A.D., Etchin,J., Lawton,L., Sallan,S.E., Silverman,L.B. *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, N.Y.)*, **346**, 1373–1377.
17. Saint-Andre,V., Federation,A.J., Lin,C.Y., Abraham,B.J., Reddy,J., Lee,T.I., Bradner,J.E. and Young,R.A. (2016) Models of human core transcriptional regulatory circuitries. *Genome Res.*, **26**, 385–396.
18. Gryder,B.E., Yohe,M.E., Chou,H.C., Zhang,X., Marques,J., Wachtel,M., Schaefer,B., Sen,N., Song,Y., Gualtieri,A. *et al.* (2017) PAX3-FOXO1 establishes myogenic super enhancers and confers BET bromodomain vulnerability. *Cancer Discov.*, **7**, 884–899.
19. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
20. Loven,J., Hoke,H.A., Lin,C.Y., Lau,A., Orlando,D.A., Vakoc,C.R., Bradner,J.E., Lee,T.I. and Young,R.A. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320–334.
21. Pott,S. and Lieb,J.D. (2015) What are super-enhancers? *Nat. Genet.*, **47**, 8–12.
22. Aldiri,I., Xu,B., Wang,L., Chen,X., Hiler,D., Griffiths,L., Valentine,M., Shirinifard,A., Thiagarajan,S., Sablauer,A. *et al.* (2017) The dynamic epigenetic landscape of the retina during development, reprogramming, and tumorigenesis. *Neuron*, **94**, 550–568.
23. Fournier,M., Bourriquen,G., Lamaze,F.C., Cote,M.C., Fournier,E., Joly-Beauparlant,C., Caron,V., Gobeil,S., Droit,A. and Bilodeau,S. (2016) FOXA and master transcription factors recruit Mediator and Cohesin to the core transcriptional regulatory circuitry of cancer cells. *Scientific Rep.*, **6**, 34962.
24. Higgins,G.A., Georgoff,P., Nikolian,V., Allyn-Feuer,A., Pauls,B., Higgins,R., Athey,B.D. and Alam,H.E. (2017) Network reconstruction reveals that valproic acid activates neurogenic transcriptional programs in adult brain following traumatic injury. *Pharmaceut. Res.*, **34**, 1658–1672.
25. Lin,C.Y., Erkek,S., Tong,Y., Yin,L., Federation,A.J., Zapatka,M., Haldipur,P., Kawauchi,D., Risch,T., Warnatz,H.J. *et al.* (2016) Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature*, **530**, 57–62.
26. Winter,G.E., Mayer,A., Buckley,D.L., Erb,M.A., Roderick,J.E., Vittori,S., Reyes,J.M., di Iulio,J., Souza,A., Ott,C.J. *et al.* (2017) BET bromodomain proteins function as master transcription elongation factors independent of CDK9 recruitment. *Mol. Cell*, **67**, 5–18.
27. Khan,A. and Zhang,X. (2016) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, **44**, D164–D171.
28. Wei,Y., Zhang,S., Shang,S., Zhang,B., Li,S., Wang,X., Wang,F., Su,J., Wu,Q., Liu,H. *et al.* (2016) SEA: a super-enhancer archive. *Nucleic Acids Res.*, **44**, D172–D179.
29. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
30. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
33. Schaefer,U., Schmeier,S. and Bajic,V.B. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, **39**, D106–D110.
34. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
35. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
36. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, **27**, 1017–1018.
37. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
38. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.
39. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–1051.
40. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
41. Zhang,S.J., Liu,C.J., Shi,M., Kong,L., Chen,J.Y., Zhou,W.Z., Zhu,X., Yu,P., Wang,J., Yang,X. *et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res.*, **41**, D892–D905.
42. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
43. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
44. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.